The genus of regular languages

Guillaume Bonfante[1], Florian Deloup[2]

[1] Université de Lorraine, LORIA, Nancy, France
[2] Université Paul Sabatier, IMT, Toulouse, France

Table of Contents

Abstract. The article defines and studies the genus of finite state deterministic automata (FSA) and regular languages. Indeed, a FSA can be seen as a graph for which the notion of genus arises. At the same time, a FSA has a semantics via its underlying language. It is then natural to make a connection between the languages and the notion of genus. After we introduce and justify the the notion of the genus for regular languages, the following questions are addressed. First, depending on the size of the alphabet, we provide upper and lower bounds on the genus of regular languages : we show that under a relatively generic condition on the alphabet and the geometry of the automata, the genus grows at least linearly in terms of the size of the automata. Second, we show that the topological cost of the powerset determinization procedure is exponential. Third, we prove that the notion of minimization is orthogonal to the notion of genus. Fourth, we build regular languages of arbitrary large genus: the notion of genus defines a proper hierarchy of regular languages.

Beyond the set-theoretic description of graphs, there is the notion of an embedding of a graph in a surface. Intuitively speaking, an embedding

of a graph in a surface is a drawing without edge-crossings. Planar graphs are drawn on the sphere S_0, the graphs K_5 and $K_{3,3}$ are drawn on the torus S_1 and more generally, any graph can be drawn on some closed orientable surface S_k, that is a sphere with k "handles". The genus of a graph G is the minimal index k such that G can be drawn on S_k.

The aim of this work is to explore standard notions of finite state automata (FSA) theory with this topological point of view. The novelty of this point of view lies in the fact that finite state automata are not only *graphs*, they are *machines*. These machines compute regular languages. The correspondence is onto: one language may be computed by infinitely many automata. It is then natural to define the genus of a regular language to be the minimal genus of its representing deterministic automata.

It should be noted that the word "deterministic" in the previous sentence is crucial: any regular language is recognized by some *planar* nondeterministic automaton. The earliest reference for this result we could find is [BoCh76]. The cost in terms in extra states and transitions is analyzed in [BP99]. By contrast, we show in this paper the existence of regular languages having *arbitrary high genus*.

The use of topology in the study of languages may come as a surprise at first. We suggest two motivations of very different nature. First, the question arises naturally if one wants to build physically the FSA. Think of boolean circuits, they also are graph-machines. There is an immense literature about their electronic implementation, that is about the layout of Very-Large Scale Integration (VLSI) (for instance [CKC83]). In particular, the problem of minimization of via is close to the current one. Many contributions suppose a fixed number of layers (holes), but some consider an arbitrary one [SHL90]. As we will show, a smaller number of states may not necessarily mean a smaller cost in terms of the electronic implementation.

There is a second and more fundamental reason why one should consider topology in general and the genus in particular in the study of regular languages. Low-dimensional topology is a natural tool in order to estimate the complexity of languages (or the complexity of the computation of languages). The main invariant of a regular language L is usually the number of states (the size) of the minimal automaton recognizing L. This invariant describes the size of the table data in which transitions are stored, that is the size of the machine's memory. However, simple

counting costs memory without complexifying the internal structure of automata. As a simple example, the language $L_n = \{a^n\}$ is represented by an automaton of size $n + 2$ but with the simple shape of a line:

The genus, as a complexity measure, has been introduced for formal logical proofs by R. Statman [Sta74], and further studied by A. Carbone [Car09]. Cut-elimination is presented as a way of diminishing the complexity of proofs, that is of simplifying proofs. We are not aware of other use of low dimensional topology as a complexity measure besides this work. To the best of our knowledge, classical textbooks (e.g. [HMU06], [Sak09], [RS97]) about automata theory are devoted to the set-theoretic approach. Our long term objective is a topological study of the well known constructions such as minimization, determinization, union, concatenation, and so on. This paper is devoted to the notion of genus.

As a first step, we derive a closed formula for the genus of a deterministic finite automaton (Theorem 5). Then we show that under a rather mild hypotheses on the size of the alphabet (≥ 4) and on the geometry, the genus of a deterministic finite automaton at least increases linearly in terms of the number of states (Theorem 6). Since the hypotheses depend only on the abstract representing automaton and not on a particular embedding, we deduce an estimation of the genus of regular languages (Theorem 7).

Theorem 1. *Let $(L_n)_{n \geq 1}$ be a sequence of regular languages L_n of size n, with alphabet size $m \geq 4$. Assume that for any deterministic automaton recognizing L_n, the number of cycles of length 1 and 2 is negligible with respect to n. For any $\varepsilon > 0$, there is $N > 0$ such that for all $n \geq N$,*

$$1 + \left(\frac{m-3}{6m} - \varepsilon \right) mn \leq g(L_n) \leq mn.$$

We present several remarkable consequences of this result throughout this paper.

We mention two particular cases of interest. It is known that the size of the union of two automata increases linearly with the product of their respective size. We prove that the *genus* of the union of two automata A and B increases linearly with the product of the sizes of A and B (Corollary 4). We also provide an example of a nondeterministic automaton A such that the genus of the powerset-determinized form of a A is exponential up to a linear factor with respect to the size of A (Theorem 8).

In a second step, we study further the link between languages, their representation in terms of automata and their genus. The comparison with state minimization is instructive. Myhill-Nerode Theorem ensures that two deterministic automata with same minimal number of states that recognize the same laguage must be isomorphic. We show that this uniqueness property does not hold if we replace minimal number of states by minimal genus. There is no simple analog to Myhill-Nerode Theorem. As a consequence nonisomorphic automata representing the same language may have minimal genus. There may be even nonisomorphic automata of minimal size within the set of genus-minimal automata.

As a final step, we describe *explicit* languages having arbitrary high genus (Theorem 9). These results imply the existence of a nontrivial hierarchy of regular languages based on the genus and yields a far-reaching generalization of the results of [BoCh76, §4]. In particular, the genus yields a nontrivial measure of complexity of regular languages.

1   Finite State Automata

We briefly recall the main definitions of the theory of finite state automata and regular languages. An *alphabet* is a (finite) set of *letters*. A word on an alphabet A is a finite sequence of letters in the alphabet. Let A^* be the set of all words on A, ϵ is the empty word and the concatenation of two words w and w' is denoted by $w \cdot w'$. We define repetitions as follows. Given some word w, let $w^0 = \epsilon$ and $w^{n+1} = w^n \cdot w$.

A language on an alphabet A is a subset of A^*. Given two languages, let $L + L'$, $L \cdot L'$ and L^* denote respectively the union, the catenation and the star-operation on L (and L'). Rational languages are those languages build from finite sets and the three former operations.

A (finite state) automaton is a 5-tuple $\mathtt{A} = \langle Q, A, q_0, F, \delta \rangle$ with Q, a finite set of *states* among which q_0 is the *initial* state, $F \subseteq Q$ is the set of *final* states, A is an alphabet and $\delta \subseteq Q \times A \times Q$ is the *transition relation*. The relation δ extends to words by setting $\delta(q, \epsilon, q)$ for all $q \in Q$ and by defining $\delta(q, a \cdot w, q')$ if and only if $\delta(q, a, q'')$ and $\delta(q'', w, q')$ for some state $q'' \in Q$. Such an automaton induces a language

$$\mathcal{L}_{\mathtt{A}} = \{w \in A^* \mid \delta(q_0, w, q_f) \wedge q_f \in F\}.$$

The language $\mathcal{L}_{\mathtt{A}}$ is said to be recognized (or represented) by $\mathtt{A}$. A fundamental result is Kleene's theorem.

Theorem 2 (Kleene). *A language is regular if and only if it is recognized by some finite state automaton.*

Example 1. On the alphabet $\{a, b\}$, let us define the automaton F on the left and K_5 on the right:

The small arrows indicate the initial states and final states are doubly circled. The language recognized by F is $\mathcal{L}_F = \{a^n \cdot b \mid n \in \mathbb{N}\} \cup \{(a \cdot b)^n \mid n > 0\} = a^* \cdot b + (a \cdot b)^*$. The language recognized by K_5 is $\mathcal{L}_{K_5}$ composed of the words of "weight" 0 modulo 5. The weight of a being 1 and the one of b being 2.

An automaton $A = \langle Q, A, q_0, F, \delta \rangle$ is said to be *deterministic* (resp. *complete*) if for any state $q \in Q$ and any symbol $a \in A$, the cardinality of the set $\{q' \in Q \mid \delta(q, a, q')\}$ is at most one (resp. at least one). In the case when A is deterministic and complete, δ is actually a function $Q \times A \to Q$. In that case, $\delta(q, u) = q'$ stands for $\delta(q, u, q')$. It is well known that regular languages restrict to (complete) deterministic automata. All deterministic automata in this paper shall be finite and complete unless stated otherwise.

Example 2. The automaton K_5 is deterministic, but F is not. Nevertheless, $\mathcal{L}_F$ is recognized by the automaton F':
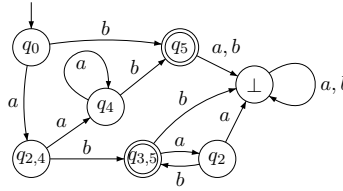
Note that the only function of the state symbolized by $\perp$ is to make the automaton F' complete. It is traditionally denoted the "trash state". Once this state is reached, the final states are unreachable.

Given a language L, a distinguishing extension of two words u and v is a word w such that $u \cdot w \in L$ and $v \cdot w \notin L$. Let R_L be the (equivalence) relation $u \, R_L \, v$ if and only if u and v have no distinguishing extension.

Theorem 3 (Myhill-Nerode [Myh57,Ner58]). *A language L is regular if and only if R_L has finitely many equivalence classes.*

Actually, the equivalence classes are the states of an automaton–called the minimal automaton– which, remarkably, is *the* smallest deterministic automaton recognizing L. By smallest, we mean the one with the minimal number of states. Thus, the notion of size of an automaton A, denoted $|A|$ in the sequel, is the number of states of A. We emphasized the determinant 'the' in the first sentence of the paragraph to stress the fact that there is only one (up to isomorphism) automaton of minimal size representing L.

Example 3. The automaton F' is minimal.

The size of an automaton serves as an evaluation of its complexity (see for instance [Yu00]). Due to Myhill-Nerode, one may define the complexity of a regular language to be the size of its minimal automaton.

Example 4. There are regular languages of arbitrary large complexity. For instance, on the alphabet $A = \{a\}$, consider for all $n > 0$ the language $L_n = \{a^n\}$ that consists of all words on A of length n. The linear automaton depicted in the introduction is the minimal automaton representing L_n: it has size $n + 2$.

Proposition 1. *If two words u and v have a distinguishing extension for a regular language L, then, for any deterministic automaton A representing L, $\delta(q_0, u) \neq \delta(q_0, v)$.*

Proof. Ad absurdum. Suppose that $\delta(q_0, u) = q_1 = \delta(q_0, v)$. Then, $\delta(q_0, u \cdot w) = \delta(q_1, w) = \delta(q_0, v \cdot w)$. Since $u \cdot w \in L$, $\delta(q_1, w) \in F$. Thus, $v \cdot w \in L$, in contradiction with the hypothesis. ∎

2 The genus of a regular language

Let A be a finite automaton. In the constructions to follow, we regard A as a graph where the vertices are the states and the edges are the transitions[3]. We simply forget about the extra structures on it (namely, the orientation and the labels of the edges). We are interested in a class of embeddings of A into oriented surfaces. Recall that a *2-cell* is a topological two-dimensional disc. An automaton is *planar* if it embeds into a 2-cell (or equivalently a sphere or a plane).

By means of elementary operations, one can show that A embeds into a closed oriented surface Σ. Among all embeddings that share that property, choose one such that the complement of the image of A in Σ is a

---

[3] In particular, two vertices may be joined by several edges.

disjoint union of a finite number of open 2-cells. Such an embedding will be called a *cellular embedding*. Again by elementary operations, one can show that there exists a cellular embedding of A.

As a very simple example, the automaton A that consists of one state and one loop embeds in the obvious fashion into the 2-sphere. Note that the geometrical realization of A coincides with the loop. The embedding is cellular because the complement of the loop is the union of two 2-cells. The same automaton embeds also into the torus T as depicted. In this case the embedding is not cellular because $T \setminus$ A is a cylinder and not a disjoint union of 2-cells.
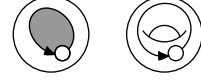
Example 5. Another example is given by an automaton with one state and two loops. Of the two embeddings depicted into the torus T, the top one is noncellular and the bottom one is cellular. (One should identify the opposite sides of the square on the left side to obtain the embedding depicted on the right side.)
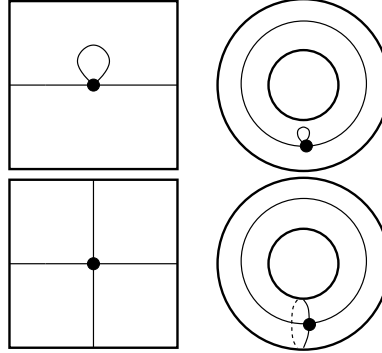
In this context, the following observation is a tautology:

Lemma 1. *A cellular embedding of an automaton* A $\subset \Sigma$ *determines a finite CW-complex decomposition of the surface Σ in which the 1-skeleton Σ^1 of Σ is the image of* A.

A CW-complex is a topological space made up of k-dimensional cells. Here we use 0-cells (points, corresponding to states), 1-cells (topological segments, corresponding to transitions) and 2-cells (topological discs). For the precise definition of a CW-complex decomposition, see for instance [Bre93, Chap. IV, §8]. For instance, the cellular embedding of A into the torus T of Example 5 induces one CW-complex decomposition of the torus consists of one 0-cell (induced by the unique state of A), two 1-cells

(induced by the two transitions of A) and one 2-cell (thought of as the complement of A in T).

Recall that the *genus* of a closed oriented surface Σ is the integer $g = \frac{1}{2}\dim H_1(\Sigma;\mathbb{R})$. In our context, it is useful to note that the genus of Σ is the maximal number of disjoint cycles that can be removed from Σ such that the complement remains connected.

Definition 1. *A cellular embedding of A into Σ is* minimal *if the genus of Σ is minimal among all possible surfaces Σ into which A embeds cellularly.*

Example 6. The second embedding of Example 5 is cellular: the complement of A consists in one open 2-cell. It is not minimal. Indeed, the automaton embeds into the 2-sphere S^2: it is realized as the wedge of two circles ◯◯ (whose complement in S^2 consists of three open 2-cells).

Definition 2. *The* genus $g(A)$ *of a finite deterministic automaton A is the genus of Σ where Σ is a closed oriented surface into which A embeds minimally.*

Example 7. The genus of the automaton that consists in one state and an arbitrary number of loops is zero because it embeds into the 2-sphere.

Let g_A be the smallest number $g_\Sigma \in \mathbb{N}$ where Σ is a closed oriented surface into which A can be embedded. Then $g_A \leq g(A)$ (since all possible embeddings, included noncellular ones, are considered).

Theorem 4 (J.W.T. Youngs [You63]). *For any automaton A, $g_A = g(A)$. In other words, an embedding with minimal genus is cellular.*

We shall use Youngs' result throughout this paper.

Example 8. Consider the example of the graph K_5, the complete graph on five vertices. It is well known that K_5 is not planar. Embed it into the torus T as depicted in Fig. 1. Since the torus has genus 1, the embedding is minimal. One verifies that it is also cellular: the complement of K_5 in T consists of five disjoint open 2-cells.

We can now formally state the definition of the genus of a regular language.

Definition 3. *Let L be a regular language. The* genus $g(L)$ *of L is the minimal genus of a complete finite deterministic automaton recognizing L:*

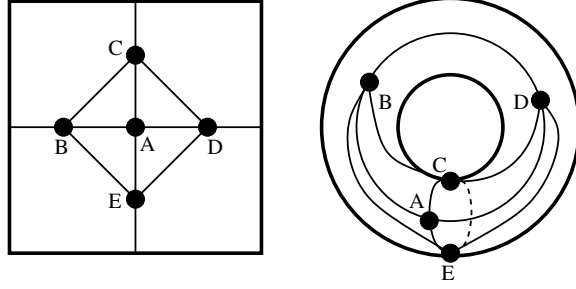$$g(L) = \min\{g(A) \mid L = L_A,\ A\ \text{complete finite deterministic}\}.$$

Fig. 1. A cellular embedding of the graph K_5 in the torus T.

There is a simple upper bound for the genus of a deterministic automaton.

Proposition 2. *Let* A *be a deterministic automaton with m letters and n states. Then*

$$g(\mathtt{A}) \leq mn.$$

Proof. This follows from Euler's formula (7). ∎

Given some fixed alphabet, Prop. 2 shows that the genus of a regular language L is smaller than the size of a minimal automaton recognizing L up to some linear factor. Hence the main problem we face is to compute a lower bound for the genus (see Th. 7).

The next two results deal with the completeness of automata and reachable states. They are instrumental in nature: they say that the completion of a automaton of minimal genus and the suppression of its unreachable states do not modify the genus. These facts will be used in the sequel without further notice.

Proposition 3. *For any regular language L with genus g, there is a complete, deterministic automaton of genus g representing L.*

Proof. Let L be a regular language with genus g. Then, there is a deterministic automaton $\mathtt{A} = \langle Q, A, q_0, F, \delta \rangle$ representing L that embeds cellularly in a surface Σ of genus $g(\mathtt{A}) = g$. First, to any state q of $\mathtt{A}$ which would not be complete, add a new trash state $\perp_q$ with the transitions $\delta(q, a) = \perp_q$ for all letter a such that $\delta(q, a)$ is not defined. Second, to each of these new trash states, add loops $\delta(\perp_q, a) = \perp_q$ for all $a \in A$. Clearly, the new transitions embed into Σ and do not modify the genus of $\mathtt{A}$. ∎

A state q of a (deterministic) automaton $\mathtt{A} = \langle Q, A, q_0, F, \delta \rangle$ is said to be *reachable* if there is a word w such that $\delta(q_0, w) = q$.

Proposition 4. *For any regular language L with genus g, there is a deterministic, complete automaton $\mathtt{A}$ of genus g representing L such that all states of $\mathtt{A}$ are reachable.*

Proof. Consider an automaton $\mathtt{A}$ of genus g representing L. Remove all unreachable states and the corresponding transitions from $\mathtt{A}$. The language recognized by the modified automaton is still L. Being a subgraph of $\mathtt{A}$, the new automaton has a genus smaller or equal to g. Since it represents the same language L, its genus must be equal to g. All its states are reachable. ∎

2.1 Combinatorial cycles and faces

In this paragraph, we introduce cycles and faces. A cycle is a notion that depends only on the abstract graph, while a face depends on a cellular embedding of the graph. The notion of faces is crucial in the Genus Formula (Theorem 5) and instrumental in the Genus Growth Theorem (Theorem 6).

Definition 4. *Let $p \geq 1$. A* walk *in $\mathtt{A}$ is a finite alternating sequence of vertices (states) and edges $s_0, t_1, s_1, t_2 \ldots, t_p, s_p$ of $\mathtt{A}$ such that for each $j = 1, \ldots, p$, the states s_{j-1} and s_j are the endpoints of the edge t_j. The* length *of the walk is the number of edges (counting repetitions). An* internal vertex *of the walk is any vertex in the walk, distinct from the first vertex s_0 and the last vertex s_p. The walk is* closed *if the first vertex is the last vertex, $s_0 = s_p$.*

Recall that we regard $\mathtt{A}$ as an unoriented graph: one can walk along an edge opposite to the original orientation of the transition. The edge should be nonempty: there should be an actual transition in one direction or the other. In particular, if there is no transition from a state s to itself, then the vertex s cannot be repeated in the sequence defining a walk.

If the underlying graph is simple, then we suppress the notation of the edges: a walk is represented by a sequence of vertices $s_0, s_1, \ldots, s_p$ such that any two consecutive vertices are adjacent.

Definition 5. *Consider the set $W(p)$ of closed walks of length p in $\mathtt{A}$. The group of cyclic permutations of $\{1, \ldots, p\}$ acts on $W(p)$. A* combinatorial cycle *of length p, or simply a p-cycle, is an orbit of a closed walk of length p.*

In other words, two closed walks represent the same combinatorial cycle if there is a cyclic permutation that sends one onto the other. This definition is propped by the fact that we are interested in geometric cycles only and we do not want to count them with multiplicities with respect with the start of each node.

Remark 1. Our definition of a cycle departs from the traditional one in graph theory: repetitions of edges and internal vertices may occur. A combinatorial cycle in which no edge occurs more than one will be called a *simple* cycle. (We still allow repetition of an internal vertex that has several loops.)

We denote the set of all p-cycles in $\mathtt{A}$ by $Z_p(\mathtt{A})$. Since $\mathtt{A}$ is finite, $Z_p(\mathtt{A})$ is finite. We set $z_p = |Z_p(\mathtt{A})|$.

The definitions of walks and cycles are intrinsic to the graph: they do not depend on an embedding (or a geometric realization) of the graph. However, they are directly related to topology once an embedding is given. Let $\mathtt{A}$ be an automaton embedded in a surface Σ. Each combinatorial cycle determines a geometric 1-cycle (in the sense of singular homology) in Σ. Therefore combinatorial loops are thought of as combinatorial analogues of singular 1-cycles (in the sense of singular homology).

In what follows, consider a cellular embedding of $\mathtt{A}$ into a closed oriented surface Σ. By definition, the set $\pi_0(\Sigma - \mathtt{A})$ of connected components of $\Sigma - \mathtt{A}$ consists of a finite number of 2-cells. The image in Σ of the set $\mathtt{A}^1$ of edges of $\mathtt{A}$ is the 1-skeleton Σ^1 of Σ. With a slight abuse of notation, we shall denote by the same symbol Σ^1 the collection of embedded edges of $\mathtt{A}$. Consider an edge $e \in \Sigma^1$ and an open 2-cell $c \in \pi_0(\Sigma - \mathtt{A})$. It follows from definitions that if $\mathrm{Int}(e)$ and $\mathrm{Fr}(c)$ intersect nontrivially then $e \subset \mathrm{Fr}(c)$. Since Σ is a 2-manifold, there is at most one component c' of $\Sigma - \mathtt{A}$, $c' \neq c$, such that $e \subset \mathrm{Fr}(c)$.

Without loss of generality, we may assume that the embedded edge e is a smooth arc. Let x be a point in e. Define a small nonzero normal vector $\overrightarrow{n}$ at x. If $\overrightarrow{n}$ and $-\overrightarrow{n}$ point to distinct components c, c' of $\Sigma - \mathtt{A}$, then $e \subset \mathrm{Fr}(c) \cap \mathrm{Fr}(c')$: there are two distinct components separated by e. In this case we say that e is *bifacial*. If $\overrightarrow{n}$ and $-\overrightarrow{n}$ point to the same component c of $\Sigma - \mathtt{A}$, then c is the unique component of $\Sigma - \mathtt{A}$ such that $e \subset \mathrm{Fr}(c)$. In this case, one says that the edge e is *monofacial*.

We define a pairing $\langle -, - \rangle : \Sigma^1 \times \pi_0(\Sigma - \mathtt{A}) \to \{0, 1, 2\}$ by

$$\langle e, c \rangle = \begin{cases} 0 \text{ if } e \cap \mathrm{Fr}(c) = \varnothing \\ 1 \text{ if } e \subset \mathrm{Fr}(c) \text{ and } e \subset \mathrm{Fr}(c') \text{ for } c' \in \pi_0(\Sigma - \mathtt{A}) - \{c\} \\ 2 \text{ if } c \text{ is the unique component of } \Sigma - \mathtt{A} \text{ such that } e \subset \mathrm{Fr}(c). \end{cases}$$

From the discussion above, it follows that for any edge $e \in \Sigma^1$,

$$\sum_{c \in \pi_0(\Sigma - \mathtt{A})} \langle e, c \rangle = 2. \tag{1}$$

Definition 6. *Let $k \geq 1$. A component c of $\Sigma - \mathtt{A}$ is a k-face if*

$$\sum_{e \in \Sigma^1} \langle e, c \rangle = k.$$

The set of k-faces is denoted F_k. We let f_k denote the number of elements in F_k. A face *is a k-face for some $k \geq 1$. The set of faces is denoted F.*

The following properties hold:

(1) The sets $F_k, k \geq 1$ are disjoint;
(2) All sets F_k but finitely many are empty;
(3) $F = \bigcup_{k \geq 1} F_k = \pi_0(\Sigma - \mathtt{A})$.

Definition 7. *Let $k \geq 1$. A combinatorial k-gon in Σ is a k-cycle of $\mathtt{A}$ that bounds a k-face of $\Sigma - \mathtt{A}$. A 2-gon will also be called a* bigon. *A cycle of length 1 will be called a* loop.

Lemma 2. *Any 1-gon has a bifacial edge.*

The proof follows from the more general fact that a contractible simple closed curve is separating. See §8.1 for a proof.

The automaton depicted opposite has two states; each state has three outgoing transitions. It is cellularly embedded into the plane. All edges are bifacial. The edge α is a loop contractible in Σ but is not a 1-gon; $\alpha\beta$ is a cycle of length 2 that is a bigon; $\gamma\delta$ is a cycle of length 2, contractible in Σ, that is not a bigon.
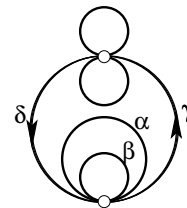
According to Lemma 2, a cycle of length 1 is monofacial if and only if it is not a 1-gon (if and only if it represents a nontrivial element in 1-homology). A bigon may have monofacial edges, even in a cellular embedding: the cellular embedding of Example 5 provides such an instance.

Remark 2. For any $k \geq 1$, $f_k \leq z_k$. Indeed, just as in homology, a combinatorial cycle does not necessarily bound a combinatorial face. For instance, in the K_5 embedded in the torus as in Example 8, the simple cycle $BCDB$ of length 3 is not a 3-gon. In fact, it does not bound any 2-cell.

Lemma 3 (1-gon lemma)**.** *There exists a minimal embedding of* A *such that any cycle of length one is a 1-gon and in particular, is bifacial.*

See §8 for a proof. In the sequel, we shall frequently use Lemma 3 without further notice.

By contrast, there exists automata for which there is no embedding such that every cycle of length two is a bigon. A simple example can be constructed using the subgraph opposite.

Our definition of a combinatorial cycle mimics that of a geometric 1-cycle c in the sense of a singular 1-chain such that $\partial c = 0$. We remark that in order to represent a singular 1-cycle by a combinatorial cycle, the combinatorial cycle in question may have repetitions of edges and internal vertices.

For instance, consider anew the cellular embedding of Example 8. As mentioned, the complement of K_5 in T has five components which are open 2-cells. Four of them are 3-faces bounded respectively by $ABCA$, $ACDA$, $ADEA$ and $AEBA$. The fifth component is an open 2-cell: removing the other four open 2-cells and the edges BD and CE yields a 2-cell. (Removing the four open 2-cells yields a punctured torus, which is a regular neighborhood of the wedge of a meridian and a longitude; removing the edges BD and CE amounts to cutting transversally the meridian and the longitude respectively, yielding a topological 2-cell.)
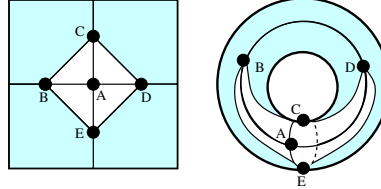
This fifth 2-cell is a bit more complicated to describe: it is not bounded by any simple cycle. It is bounded by the closed walk $BCEBDECDB$ which represents a combinatorial cycle of length 8. It therefore represents an 8-gon. Note that the monofacial edges CE and BD are travelled twice in opposite orientations.
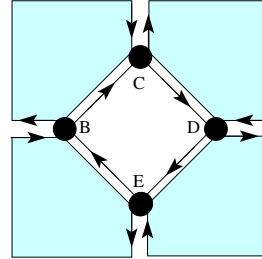
2.2   Digression: face embeddings and strong face embeddings

This paragraph is not necessary to understand our results and their proofs (and hence may be skipped on a first reading). Indeed they do not depend on the notions introduced here. In particular, they do not depend on whether the Strong Embedding Conjecture (or a related conjecture) is true or not. It is true however that for a graph that has a strong embedding in a surface of minimal genus, then the Genus Formula has a particularly simple form. (However, it is known in general that this needs not always be the case. There exist 2-connected graphs of genus 1 that have no strong cellular embedding in a torus, see [Xuo77].) We include this paragraph for clarification.

As we have seen, k-gons that appear in cellular embeddings need not be simple. We may request them to be, at the expense of a more restrictive definition.

Definition 8. *A* face embedding *of a graph* A *into a closed oriented surface* Σ *is a cellular embedding of* A *into* Σ *such that each k-face in $\Sigma - $A is bounded by a simple k-gon.*

Opposite is depicted another embedding of K_5 into the torus (the opposite sides in the square are identified as usual). This embedding is a face embedding of K_5: the complement of K_5 consists of five 4-faces. Hence this embedding is not equivalent to the cellular embedding of Example 8.
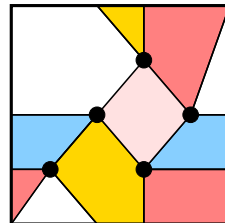
Note that a face embedding does not rule out the possibility that an edge be monofacial. Recall that e is bifacial if there is a component c in $\Sigma - $A such that $\langle e, c \rangle = 1$.

Definition 9. *Let e be an edge of embedded graph* A *into a closed oriented surface* Σ. *A* strong face embedding *of a graph* A *into a closed oriented surface* Σ *is a face embedding of* A *into* Σ *such that every* edge *is bifacial.*

For instance, in Example 8, all but the edges BD and CE are bifacial. The face embedding of K_5 above is strong. The second embedding of Example 5 is a face embedding that is not strong.

This definition is related to that of strong cellular embedding: a strong cellular embedding of A is an embedding in Σ such that the closure of each connected component $\Sigma - $A is a closed 2-cell. Equivalently, every k-face of $\Sigma - $A is bounded by a true simple cycle without repetition of an internal vertex. A strong cellular embedding is a strong face embedding. The converse does not hold in general.

The strong cellular embedding conjecture [Jae85] is that every 2-connected graph has a strong cellular embedding into some closed surface (orientable or not). Even though the question is theoretically simpler, we do not know whether every 2-connected graph has a strong face embedding into some closed surface (orientable or not).

3 Genus Formula

Our first main result is a closed formula for the genus of a regular language.

Theorem 5 (Genus formula). *Let* A *be a deterministic automaton with* m *letters. Then for any cellular embedding of* A,

$$g(\mathtt{A}) \leq 1 - \frac{m+1}{4m}f_1 - \frac{1}{2m}f_2 + \frac{m-3}{4m}f_3 + \frac{2m-4}{4m}f_4 + \frac{3m-5}{4m}f_5 + \cdots \quad (2)$$

with equality if and only if the embedding is minimal.

The faces $f_1, f_2, \ldots$ are determined by the cellular embedding of A. It follows from §2.1 that for each cellular embedding, there is some $M > 0$ such that $f_k = 0$ for all $k \geq M$. In particular, the sum $\sum_{k=1}^{\infty} \frac{k(m-1)-2m}{4m}f_k$ that appears on the right hand side of (2) is finite.

Remark 3. In the case when (2) is an equality, it is *not* claimed that the embedding is unique. Thus inequivalent minimal embeddings for A lead to distinct formulas for the genus of A.

Remark 4. In the case when (2) is an equality, it is *not* claimed that the automaton A is the minimal state automaton (in the sense of Myhill-Nerode). Indeed, the automaton with the least number of states does not have necessarily minimal genus (see below §5).

The Genus Formula is proved in §9.
The following corollary is a straightforward consequence.

Corollary 1. *Let L be a regular language on m letters. For any minimal embedding of a deterministic automaton* A *representing L,*

$$g(L) \leq 1 + \sum_{k=1}^{\infty} \frac{k(m-1)-2m}{4m}f_k. \qquad (3)$$

4 Genus growth

Let us begin with a simple example. Any language on a 1-letter alphabet is represented by a planar deterministic automaton. Indeed, these deterministic automata have a quasi-loop (planar) shape:
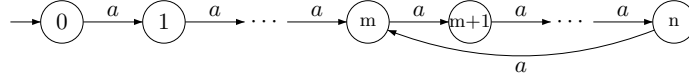
Actually, there is an indirect proof of the result. For a unary alphabet, $e_1 \leq e_0$. Since there is at least one face, Euler's relation states that $1 \leq 2 - 2g$, that is $g \leq 1/2$. And thus $g = 0$. The remark shows that to get graphs of higher genus, one should augment the number of edges. Then the relation $e_1 = me_0$ forces to augment the size of alphabets. Thus a general study of the genus of automata depends on the size of alphabet. Consider now the case of a 2-letter alphabet.

Proposition 5. *Let $(\mathtt{A}_n)_{n\in\mathbb{N}^\times}$ be a sequence of deterministic finite automata of size n on the same alphabet. Assume that the number m of letters is two. For any cellular embedding of $\mathtt{A}_n$,*

- *either there exists $M > 0$ such that $\sup\limits_{n\geq 1} \big(f_1(n) + f_2(n) + f_3(n)\big)\leq M$*
- *or $\lim\limits_{n\to+\infty} \sum_{k\geq 5} \frac{k-4}{8} f_k(n) = +\infty$.*

Proof. Suppose neither condition is satisfied. Then

$$f_1(n) \underset{n\to+\infty}{\to} +\infty \ \text{ or } \ f_2(n) \underset{n\to+\infty}{\to} +\infty \ \text{ or } \ f_3(n) \underset{n\to+\infty}{\to} +\infty$$

(since these are sequences of nonnegative integers) and $\sum_{k\geq 5} f_k(n)$ remains bounded. For $m = 2$, the second, third and fourth terms respectively in the genus inequality (2) are negative or zero. The fifth term is always zero for $m = 2$. It follows easily that $g(n)$ is negative for n large enough, which is a contradiction. ∎

Corollary 2. *Let $(L_n)_{n\in\mathbb{N}}$ be a sequence of regular languages on two letters. If for each n, L_n is recognized by a deterministic automaton $\mathtt{A}_n$ of size n having a cellular embedding such that $\sum_{k\geq 5} \frac{k-4}{8} f_k(n)$ remains bounded as $n \to +\infty$, then the genus $g(L_n)$ of L_n also remains bounded.*

Proof. By hypothesis, the number $f_k(n)$ of k faces in a cellular embedding of $\mathtt{A}_n$ into a surface Σ_n verify $\sum_{k\geq 5} \frac{k-4}{8} f_k(n) < +\infty$. By Prop. 5, $f_1(n)$, $f_2(n)$ and $f_3(n)$ are bounded. The genus formula then shows that $g(\mathtt{A}_n)$ remains bounded. Since $g(L_n) \leq g(\mathtt{A}_n)$, the conclusion follows. ∎

Each A_n in general has several nonequivalent cellular embeddings. But for *any* cellular embedding, the alternative of Prop. 5 holds for the various numbers of faces $f_k(n)$ (determined by the embedding). Corollary 2 states a sufficient condition for the genus of a *language* to be bounded. The interest in this result lies in the fact that it discriminates between the respective contribution of the faces to the genus.

There is a similar result when the number of letters is three.

Proposition 6. *Let $(A_n)_{n \in \mathbb{N}^\times}$ be a sequence of deterministic finite automata of size n on the same alphabet. Assume that the number m of letters is three. For any minimal embedding of A,*

- *either there exists $M > 0$ such that $\sup\limits_{n \geq 1} \big(f_1(n) + f_2(n)\big) \leq M$*
- *or $\lim\limits_{n \to +\infty} \sum_{k \geq 4} \frac{k-3}{6} f_k(n) = +\infty$.*

Corollary 3. *Let $(L_n)_{n \in \mathbb{N}}$ be a sequence of regular languages on two letters. If for each n, L_n is recognized by a deterministic automaton A_n of size n having a cellular embedding such that $\sum_{k \geq 4} \frac{k-3}{6} f_k(n)$ remains bounded as $n \to +\infty$, then the genus $g(L_n)$ of L_n also remains bounded.*

The proofs of Prop. 6 and Cor. 3 are similar to those of Proposition 5 and Corollary 2.

We state our main result on the genus growth of automata and languages.

Theorem 6 (Genus Growth). *Let $(A_n)_{n \in \mathbb{N}^\times}$ be a sequence of deterministic finite automata with m letters and $n \geq 1$ states. Let $g(n)$ be the genus of A_n. Assume*

(1) *$m \geq 4$.*
(2) *The numbers $z_k(n)$ of cycles of length 1 and 2 in A_n are negligible with respect to the size n of A_n: $\lim\limits_{n \to +\infty} \frac{z_1(n)}{n} = \lim\limits_{n \to +\infty} \frac{z_2(n)}{n} = 0$.*

Then for any $\varepsilon > 0$, there exists $N > 0$ such that for all $n \geq N$,

$$g(n) \geq 1 + \left(\frac{m-3}{6m} - \varepsilon \right) mn.$$

The Genus Growth Theorem is proved in §10.

We begin with examples showing that we cannot easily dispense with the hypotheses of Theorem 6.

Example 9 (Quasi-loop automaton). Any quasi-loop finite deterministic automaton with alphabet of cardinality $m \leq 1$ is planar. This of course does not contradict the Genus Growth Theorem because Hypothesis (1) does not hold.

Example 10 (Genus 1 automaton). Let $n \geq 3$. Define an automaton $\mathtt{A}_n$ as follows. Consider the set $\tilde{S}_n = \{(i,j) \mid 0 \leq i, j \leq n\}$ inside the square $C = [0, n] \times [0, n]$. The quotient T of C under the identifications $(0, t) = (1, t)$ and $(t, 0) = (t, 1)$ is a torus. The image S_n of $\tilde{S}_n$ in T is the set of states. Note that there are exactly n^2 states. For each $(i, j) \in (\mathbb{Z}/n\mathbb{Z})^2$, define two outgoing transitions $(i, j) \to (i + 1, j)$ (mod n) and $(i, j) \to (i, j+1)$ (mod n). Choosing arbitrary initial and final states yields a finite deterministic complete automaton $\mathtt{A}_n$ with n^2 states. Clearly $g(\mathtt{A}_n) \leq 1$ for any $n \geq 3$. (This also follows from Cor. 2.) This does not contradict the Genus Growth Theorem because the alphabet has only two letters. The same example with an extra outgoing transition $(i, j) \to (i+1, j+1)$ (i.e., with an extra letter for the alphabet) still yields an automaton $\mathtt{B}_n$ with $g(\mathtt{B}_n) \leq 1$ for any $n \geq 3$. (This also follows from Cor. 3.) This still does not contradict the Genus Growth Theorem: the alphabet has only three letters.

Example 11 (Another genus 1 automaton). Start with the previous example $\mathtt{B}_n$. To each state (i, j), add an outgoing transition pointing to (i, j). This yields a deterministic complete automaton with n^2 states and an alphabet that consists now of 4 letters. There is an obvious cellular (minimal) embedding in the torus T as before. This does not contradict the Genus Growth Theorem because now the number of cycles of length one (loops) is n^2 (the number of states), so the second hypothesis is not satisfied.
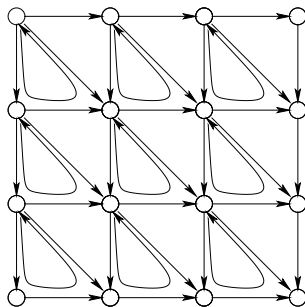
A natural consequence of the Genus Growth Theorem for automata is an estimation of the genus of regular languages.

Theorem 7 (Genus Growth of Languages). *Let $(L_n)_{n \in \mathbb{N}}$ be a sequence of regular languages on m letters, $m \geq 4$. Suppose that for each n large enough, any automaton recognizing L_n has at least n states and that the number of its cycles of length 1 and 2 are negligible with respect to n. Then for any $\varepsilon > 0$, there is $N > 0$ such that for all $n \geq N$,*

$$1 + \left(\frac{m-3}{6m} - \varepsilon \right) mn \leq g(L_n) \leq mn$$

Proof. The upper bound for the genus follows from Prop. 2. The lower bound is a direct consequence of the Genus Growth Theorem. ■

In particular, under the hypothesis of Theorem 7, the genus $g(L_n)$ grows linearly in the size n of the minimal automaton $\mathtt{A}_n$ representing L_n.

We take up the question of explicitly constructing such a sequence of regular languages in §6. There we detail an explicit construction, that shows that there is a hierarchy of regular languages based on the genus (Th. 9).

Another application of the Genus Growth Theorem is the estimation of the genus of product automata.

4.1   The genus of product automata

It is well know that the size of the product automaton corresponding to the union of two deterministic automata $\mathtt{A}$ and $\mathtt{B}$ is bounded by $m \times n$, the product of the size of $\mathtt{A}$ and the size of $\mathtt{B}$. This bound is actually a lower bound as presented by S. Yu in [Yu00]. By Prop. 2, up to a linear factor due to the size of the alphabet, $m \times n$ is also an upper bound on the genus of the product automaton. We prove that it is also a lower bound.

Corollary 4. *There is a family $(\mathtt{A}_m, \mathtt{B}_n)_{m \in \mathbb{N}, n \in \mathbb{N}}$ of planar automata $\mathtt{A}_m$ and $\mathtt{B}_n$ of respective size m and n such that the deterministic minimal automata $\mathtt{A}_n \cup \mathtt{B}_m$ has genus $O(m \times n)$.*

Proof. Let $\mathtt{A}_m$ be the m-state automaton defined as follows.
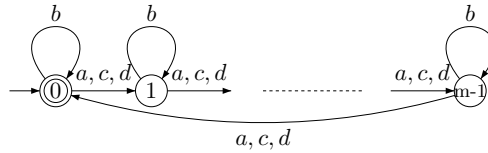
Let $\mathtt{B}_n$ be the n-state automaton defined as follows.
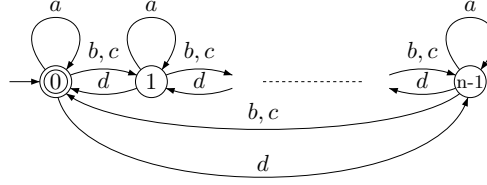
The minimal automaton $\mathtt{A}_m \cup \mathtt{B}_n$ has size $m \times n$ and it contains neither loops, nor bigons. Thus, Theorem 6 applies and leads to the conclusion. ∎

4.2 The exponential genus growth of determinization

We prove that determinization leads to an exponential growth, as this is the case for state-complexity (see for instance [GMRY12]). Consider the following family of automata $(\mathtt{A}_n)_{n \in \mathbb{N}^\times}$. The alphabet $\mathscr{A}_n = \{x_1, \ldots, x_n\}$ is a set of cardinality n. The states of $\mathtt{A}_n$ consist of one initial state s_0, n states (one state for each letter) $s_1, \ldots, s_n$ and one trash state. All states except the initial state and the trash state are final. The transitions of $\mathtt{A}_n$ are defined as follows:

- From the initial state s_0 to each state s_i ($1 \le i \le n$), there are $n-1$ transitions whose labels lie in $\mathscr{A}_n - \{x_i\}$.
- From each state s_i ($1 \le i \le n$) to itself, there are $n-1$ transitions whose labels lie in $\mathscr{A}_n - \{x_i\}$.
- From each state s_i ($1 \le i \le n$) to the trash state, there is one transition whose label is x_i. (One can add n loops with labels $x_1, \ldots, x_n$ to the trash state so that the resulting automaton is complete.)

If follows from the definition that the language recognized by $\mathtt{A}_n$ is the set of words containing at most $n-1$ distinct letters. It is also clear from the definition that for any $n \ge 2$, $\mathtt{A}_n$ is planar and nondeterministic. (Note that the fact that we include or not the trash state with or with its loops is irrelevant.)

Theorem 8. *The determinization of $\mathtt{A}_n$ is minimal and has genus*

$$g_n \ge 1 + \left(\frac{n}{4} - 1\right) 2^{n-1}.$$

For instance, $g_4 \ge 1$ so the determinization $\mathtt{A}_4^{\mathrm{det}}$ of $\mathtt{A}_4$ is not planar. This can be seen by Kuratowski's theorem (as it is can be seen $\mathtt{A}_4^{\mathrm{det}}$ contains the utility graph $K_{3,3}$). It is not hard to embed $\mathtt{A}_3^{\mathrm{det}}$ into a plane so $g_3 = 0$. Of course the meaning of the theorem is that the genus of $\mathtt{A}_n^{\mathrm{det}}$ grows at least exponentially in n.
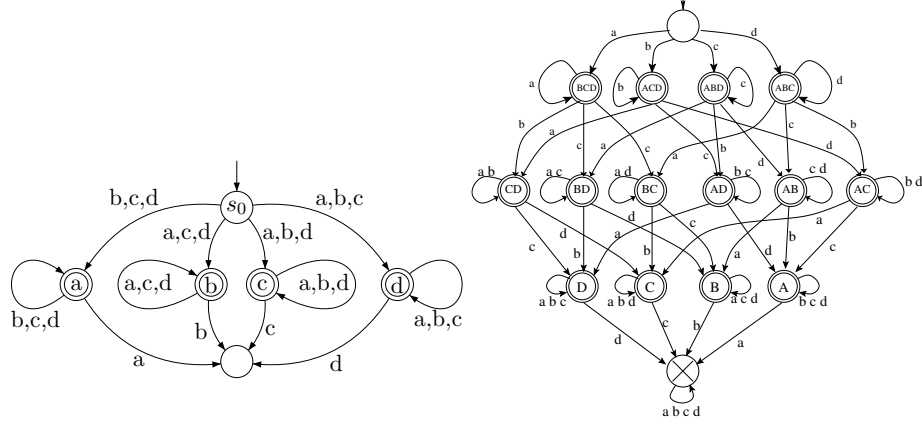
Fig. 2. The automaton A_4 and its determinized form.

Proof. Let us describe an isomorphic variant A_n^{det} of the determinized form of A by the powerset method. The states of A_n^{det} consist of all subsets of Σ_n. The initial state of A_n^{det} is $\mathscr{A}_n$ itself. The trash state is the empty set. Any state but the trash state and the initial state is a final state.

Therefore, the number e_0 of states of A_n^{det} is 2^n. The transitions are described as follows. For each letter $x \in \mathscr{A}_n$, there is one transition from S to the state $S - \{x\}$. Minimality follows from definitions: there are no indistinguishable states.

Let us consider the number e_1^o of transitions of A^{det} that are loops. By definition, each state labelled by a subset of cardinality k contributes exactly $n - k$ loops. We conclude that

$$e_1^o = \sum_{k=0}^n \binom{n}{k} (n-k) = \sum_{k=0}^n \binom{n}{k} k = n \cdot 2^{n-1}. \tag{4}$$

It follows that exactly half of the transitions are loops:

$$e_1 = 2e_1^o. \tag{5}$$

Consider now a minimal embedding of A_n^{det} into a closed oriented surface Σ. Consider one loop l in Σ^1. Since it is bifacial, it is the intersection of exactly two distinct adjacent closed 2-cells. Therefore removing the loop (while keeping the state) amounts to merging two 2-cells into one 2-cell. The union of states and transitions (minus l) still induces a CW-complex decomposition of Σ. Therefore, according to Euler's relation, the genus of Σ is unaffected. We can therefore remove all loops from Σ^1. Thus we can assume that $e_1 = e_1^o$ (from (5)) and $f_1 = 0$.

Lemma 4. *For the new graph minimally embedded in Σ, the following properties hold:*

- $f_2 = 0$;
- *For any $k \geq 1$, $f_{2k+1} = 0$;*
- *For any $k \geq n$, $f_{2k} = 0$.*

Proof. These observations are consequences of the particular structure of the original graph $\mathtt{A}_n^{\mathrm{det}}$: they follow from the definition of A^{det} and are left to the reader. ∎

We return to the proof of Theorem 8. We have

$$2e_1 = f_1 + 2f_2 + 3f_3 + 4f_4 \cdots = 4f_4 + 6f_6 + \cdots + (2n-2)f_{2n-2}.$$

The first equality is relation (9) and the second equality follows from Lemma 4. Since all numbers are nonnegative numbers, we have

$$2e_1 \geq 4(f_4 + f_6 + \cdots + f_{2n-2}) = 4 \, e_2.$$

Thus $e_2 \leq \frac{1}{2}e_1$. From Euler's relation, we deduce that

$$2g = 2 - e_0 + e_1 - e_2 \geq 2 - e_0 + e_1 - \frac{1}{2}e_1 = 2 - e_0 + \frac{1}{2}e_1.$$

Substituting values for e_0 and e_1, we obtain

$$2g \geq 2 + \left(\frac{n}{4} - 1\right) 2^n.$$

This is the desired result. ∎

5 State-minimal automata versus genus-minimal automata

Minimal automata–as given by Myhill-Nerode Theorem–have the remarkable properties to be unique up to isomorphism, leading to a fruitful relation between rational languages and automata. In this section, we show that state-minimality is a notion orthogonal to genus-minimality. First, consider the following proposition:

Proposition 7. *There are deterministic automata with a genus strictly lower than the genus of their corresponding minimal automaton.*
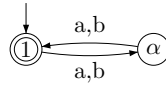
Proof. Let K_5, K' be the automata:
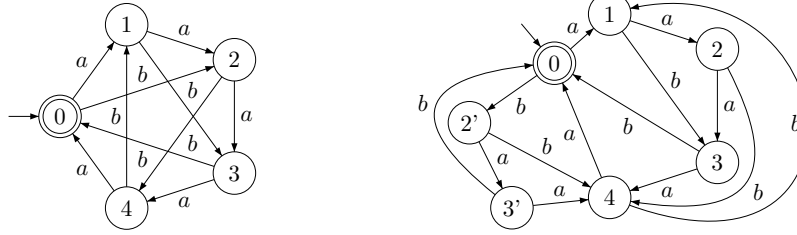
Clearly, K_5 and K' represent the same language $\mathcal{L}$, K_5 is minimal, K_5 has genus 1 and K' is planar.

Example 12. Minimal automata have not necessarily the maximal genus. For instance, the following deterministic automaton
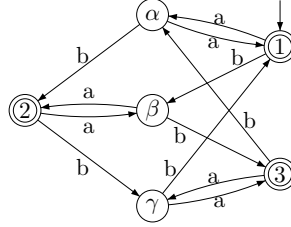

has genus 1, but its minimal corresponding automaton

has genus 0.

Contrarily to state minimization, there is no isomorphism between genus-minimal automata, even restricted to minimal state size. Indeed, the automaton K'' below on the right represents $\mathcal{L}$, but it is not isomorphic to K' on the left (the vertex 4 of automaton K' has 6 adjacent edges, the automaton K'' has no vertex with 6 adjacent edges):
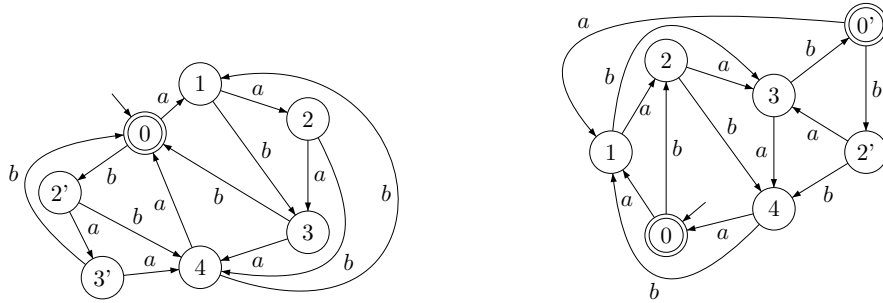
To sum up, the two automata K' and K'' (a) represent the same language, (b) have equal and minimal genus, (c) have minimal size given that genus, (d) have non isomorphic underlying graphs.

6 A hierarchy of regular languages

Is there always a planar deterministic representation of a regular language? For finite languages, the answer is positive. Indeed, finite languages are represented by trees (which are planar). In general, as evidenced by the Genus Growth Theorem for regular languages, the answer is a big "no". This section is devoted to an explicit constructive proof.

Theorem 9 (Genus-Based Hierarchy). *There are regular languages of arbitrarily large genus.*

Proof. Consider the alphabet $A = \{a, b, c, d\}$. For all $n > 0$, let U_n be the intersection of the languages $\mathcal{L}(A_n)$ and $\mathcal{L}(B_n)$ (see §4.1 for the definition). Consider an arbitrary deterministic automaton $C_n = \langle Q, A, s, F, \delta \rangle$ representing U_n.

First, as justified by Proposition 4, we can suppose without loss of generality that any state of C_n is reachable.

Second, the automaton C_n is necessarily complete. Indeed, for any word w, there is an extension of w which belongs to the language (since the final state can be reached), so that any word must be read completely.

The automaton C_n has the following key properties:

  (i) C_n has at least n^2 states,
 (ii) C_n has no loops,
(iii) C_n has no bigons.

Since the size of the alphabet A is 4, by the genus formula (2) (Th. 5), we conclude from (i–iv) that $g(C_n) \geq O(n^2)$ and the result follows. So, it remains to prove the four properties above.

Let us consider the function $\alpha : A \to \mathbb{N} \times \mathbb{N}$ with $\alpha(a) = (1, 0)$, $\alpha(b) = (0, 1)$, $\alpha(c) = (1, 1)$ and $\alpha(d) = (-1, 1)$. The function α extends to words in A^* by means of the equations $\alpha(\epsilon) = (0, 0)$ and $\alpha(e \cdot w) = \alpha(e) + \alpha(w)$ mod (n, n) for $e \in A$, $w \in A^*$ and $(x, y) + (x', y') \mod (t, u) \triangleq (x + x' \mod u, y + y' \mod v)$. A crucial observation is that
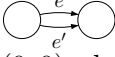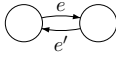
$$w \in U_n \iff \alpha(w) = (0, 0). \tag{6}$$

Remark that α induces a function–again denoted α–on states defined as follows. For the initial state s, let $\alpha(s) = (0, 0)$. Consider now a state

$q \neq s$. Since it is reachable, there is some word w such that $\delta(s, w) = q$. Define $\alpha(q) = \alpha(w)$. Note that the function is well-defined, that is, the definition does not depend on w. Indeed, suppose the existence of a word w' such that $\delta(s, w') = q$ and $\alpha(w') \neq \alpha(w)$. Let $\alpha(w) = (i, j)$. Consider $v = a^{n-i} \cdot b^{n-j}$. Then, $\alpha(w \cdot v) = (0, 0) \neq \alpha(w' \cdot v)$. And thus, v is a distinguishing extension of w and w'. This implies $\delta(s, w) \neq \delta(s, w')$ in contradiction with the hypothesis.

We come now to the three properties.

(i) Let $0 \leq i \leq n-1$ and $0 \leq j \leq n-1$. Consider the word $v_{i,j} = a^i \cdot b^j$. Then, $\alpha(v_{i,j}) = (i, j)$. Thus, there are at least $n \times n$ states.

(ii) Suppose that at state q, there is a loop labelled by a letter $e \in A$. Since q is reachable, there is some word w such that $\delta(s, w) = q$. Since $\delta(s, w \cdot e) = \delta(s, w)$, we can state that $\alpha(w \cdot e) = \alpha(w)$, which itself implies $\alpha(e) = (0, 0)$. But there is no such letter e in A.

(iii) Similarly to (ii), since there are no two letters e and e' such that $\alpha(e) = \alpha(e')$, there are no bigons of the shape: . Since there are no two letters e and e' such that $\alpha(e) + \alpha(e') = (0, 0)$, there are no bigons of the shape: . Thus, there are no bigons. ∎

7 Nondeterministic planar representation

The genus of a regular language L was defined in §2 as the minimal genus of a *deterministic* automaton recognizing L. In this section, we point out that that the word "deterministic" is crucial in the previous sentence. The following result is essentially proved by R.V. Book and A.K. Chandra [BoCh76, Th. 1a & 1b]. (See also [BP99].)

Theorem 10 (Planar Nondeterministic Representation). *For any regular language L, there exists a planar nondeterministic automaton* A *recognizing L.*

Proof. We include two proofs for the convenience of the reader. Both follow closely [BoCh76] with minor modifications. Let $L = L(R)$ be a regular language given by a regular expression R. We shall show that $L = L(A)$ for some planar nondeterministic automaton A.

The proof follows the recursive definition of a regular expression. An expression that is not the empty string is regular if and only if it is constructed from a finite alphabet using the operations of union, concatenation and Kleene's $^+$-operation. Consider the class $\mathcal{C}$ of planar finite

nondeterministic automata that have exactly one initial state, exactly one final state such that the initial state and the final state are distinct.

Clearly, $\mathcal{C}$ contains an automaton that recognizes the regular expressions $R = \varnothing$ (take A to be the automaton with two states, one initial, one final and no transition) and $R = a \in A$ (take the automaton with two states, one initial, one final and one a-labelled transition from the initial state to the final state).

Next, we show that the class $\mathcal{C}$ is closed under the three operations mentioned above. Suppose given two subexpressions R and S recognized by A and B in $\mathcal{C}$ respectively.

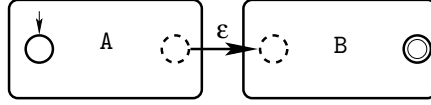Consider union: first we construct an automaton A+B with ε-transitions that recognizes $R + S$.
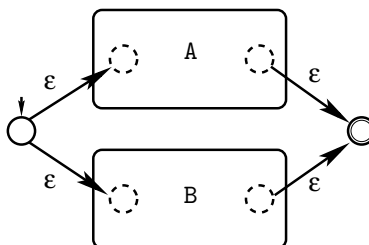
Define an ε-removal operation as follows. Consider an ε-transition that goes from state q_1 to state q_2. (We assume that $q_1 \neq q_2$.) We suppress the ε-transition and merge the two states q_1 and q_2 into one state q. Ascribe all incoming and outgoing transitions at q_1 and q_2 respectively, to the new state q. The ε-removal is best visualized by pulling the state q_2 back to the state q_1, or by pushing the state q_1 forward to the state q_2 before actually merging them[4].

---
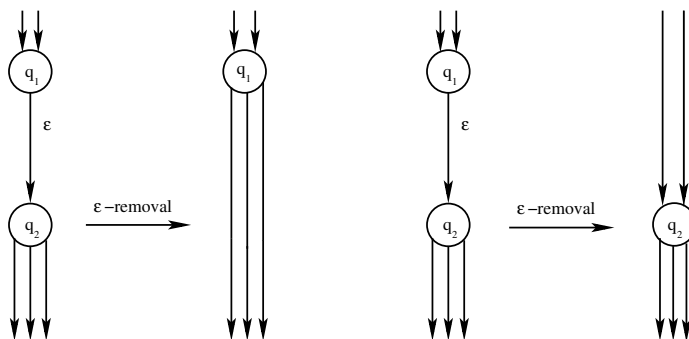
[4] Note that the result of the ε-removal operation does not depend on the orientation of the transitions.
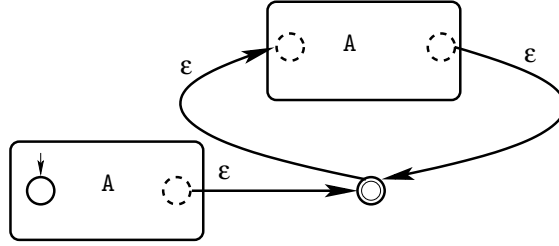
We apply this operation four times (in any order) to the automaton above. Clearly the result is an automaton that remains in $\mathcal{C}$.

Consider concatenation: the following planar automaton with one ε-transition recognizes the expression $R \cdot S$.

Next we remove the ε-transition by the ε-removal operation. This provides us with the desired automaton in $\mathcal{C}$.

Finally consider Kleene's operation: suppose that the automaton $\mathtt{A}$ recognizes the expression R. The following planar automaton with three ε-transitions recognizes $R^+ = \cup_{k \geq 1} R^k$.

We remove the ε-transitions as before. This leaves us with the desired automaton in $\mathcal{C}$. This finishes the first proof.

The second proof is short but clever. Define $\mathtt{A}_n$ be the following deterministic finite automaton with set of states $[n] = \{1, \ldots, n\}$ and alphabet $A_n = \{\sigma_{ij} \mid 1 \leq i, j \leq n\}$. For $1 \leq i, j \leq n$, set a transition with symbol σ_{ij} from i to j. We take 1 to be the initial state and 2 to be the final state.

Claim 1. *The automaton $\mathtt{A}_n$ has the universal property that any nondeterministic n-state automaton $\mathtt{A} = ([n], A, 1, \delta, 2)$ can be recovered (up to equivalence) by "parallelization" of the transitions of $\mathtt{A}_n$.*

Proof of the claim. Build an n-state automaton $\mathtt{C}_n$ by replacing each transition $\overset{\sigma_{ij}}{\textcircled{q_i} \longrightarrow \textcircled{q_j}}$ in $\mathtt{A}_n$ by $T_{ij} = \{\sigma \in A \mid j \in \delta(i, \sigma)\}$. (If T_{ij} is empty, then we remove the transition σ_{ij}. Otherwise, we have $|T_{ij}|$ distinct "parallel" transitions from i to j.) The automaton $\mathtt{C}_n$ is equivalent to $\mathtt{A}$. ∎

Claim 2. *If $\mathtt{A}_n$ has an equivalent planar automaton $\mathtt{B}_n$ then any nondeterministic n-state automaton $\mathtt{A} = ([n], A, 1, \delta, 2)$ has an equivalent planar automaton.*

Proof of the claim. We process the same proof as above with B_n instead of A_n, observing that parallelization preserves planarity.　■

It remains to construct a planar automaton B_n equivalent to A_n. The construction goes by induction on n. For $n = 3$, as a graph, A_3 is the complete graph on three vertices, hence is planar. So we take $B_3 = A_3$. Suppose we have constructed a planar automaton B_n, equivalent to A_n, together with an embedding of B_n into $\mathbb{R}^2$ and a surjective map $\alpha : Q_n \to [n]$ from the set of states of B_n to the set of states of A_n. We have to construct B_{n+1}. Consider $B_n \subset \mathbb{R}^2$. For any pair of distinct states q, q' of B_n, merge the transitions from q to q' and from q' to q into one unoriented edge. (If there is no transition, we do not perform any operation.) Finally we remove loops at each state. We obtain in this fashion an undirected simple graph G_n whose vertices are exactly the states of B_n. For each face f of $\mathbb{R}^2 - G_n$, place one vertex v inside f except for the exterior face (the unbounded component of $\mathbb{R}^2 - G_n$), and connect it to all vertices of the face f and itself. We obtain a new graph G_{n+1}. See the figure below for the recursive constructive of G_3, G_4 and G_5.
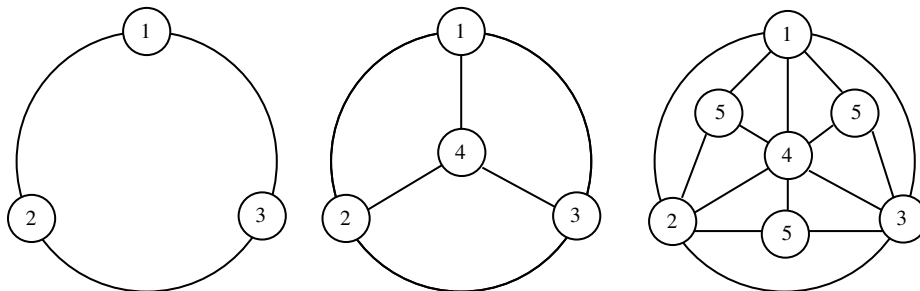
We extend α by setting $\alpha(v) = n+1$. We restore all previous (oriented) transitions between any pair of vertices, we label the new loop at v by the symbol $\sigma_{n+1,n+1}$ and we unfold each newly created edge from v to any other (old) vertex w into two transitions with opposite orientations with symbols $\sigma_{n+1,\alpha(w)}$ and $\sigma_{\alpha(w),n+1}$ respectively.
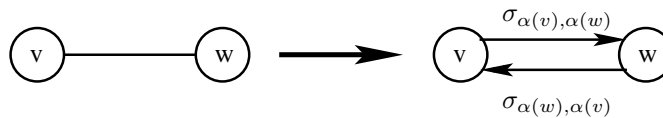
This yields a new automaton B_{n+1}. It is clear that the recursive step does not affect the initial state and the final state of B_{n+1} (that were already constructed together with B_3). The automaton B_{n+1} is planar

since G_{n+1} is planar and the unfolding of the edges preserves planarity. It remains to see that B_{n+1} is equivalent to A_{n+1}. It follows from the definition of B_n that for $q, q' \in Q_n$ and $\sigma_{ij} \in A_n$, there is a σ_{ij}-transition from q to q' if and only if $\alpha(q) = i$ and $\alpha(q') = j$. It follows that every word recognized by B_n is also recognized by A_n. To prove the converse, one shows that for any sequence $x_1 = 1, x_2, \ldots, x_k = 2$ in $[n]$ (which is a word in the language recognized by A_n), there is a path[5] $y_1, \ldots, y_k$ in B_n such that $\alpha(y_j) = j$ for each $1 \leq j \leq k$. This is proved by induction on $k \leq m$ by using the facts that it is true for $m = 3$ and that G_n contains isomorphic copies of G_{n-1}. ∎

8 Proof of the 1-gon Lemma

8.1 Proof of Lemma 2

Geometrically, a bifacial embedded loop is nothing else than a separating simple closed curve with a basepoint. It suffices to prove that a contractible simple closed curve is separating. Consider an embedded loop α in Σ^1 based at $q \in \Sigma^0$. Assume that α is monofacial (nonseparating). Consider a small segment I transversal (say, normal) to α such that $I \cap \alpha = \{q\}$. Since α is monofacial, the endpoints of I lie in the same connected component of $\Sigma - A$. Hence I extends to a loop β such that $\beta \cap (\Sigma - A) = \beta \cap \alpha = \{q\}$. It follows that the algebraic 1-homology intersection $[\beta] \cdot [\alpha] = \pm 1$. In particular, $[\alpha] \neq 0$ in $H_1(\Sigma)$. Thus α is not contractible. ∎

8.2 Proof of the 1-gon Lemma

Consider a state q of $A \subset \Sigma$ that has at least one noncontractible loop. Consider a small enough open disc D in Σ centered in q such that the following properties hold: 1) $D \cap (\Sigma - A)$ is a disjoint union of open cells; 2) The intersection $D \cap A$ is a wedge of semi-open arcs intersecting in their common endpoint q; 3) Each arc α is bifacial: there are exactly two adjacent cells $c, c' \in C = \{c_1, \ldots, c_r\}$ such that $\alpha \subset \mathrm{Fr}(c) \cap \mathrm{Fr}(c')$.

Let A be the set of arcs. The orientation of Σ induces a circular ordering $\alpha_1, c_1, \alpha_2, c_2, \ldots, \alpha_r, c_r$ of $A \cup D$ where the arcs and cells alternate and such that any two consecutive cells are adjacent.

We fix now an arc α_1 and perform successively the following operations on the arcs following the circular ordering. If the arc α_j does not belong

---

[5] A path is a walk such that no edge occurs more than once and no internal vertex is repeated.

to a loop (i.e. is part of a transition that is not a loop), we do not do anything. Otherwise there is another arc β belonging to the same loop. If the two arcs are enumerated consecutively in the circular ordering, we remove the whole loop inside Σ and replace it by a small 1-gon ℓ based at q such that $\ell - q$ lies entirely in the open cell c_j. At the end of the process, we have replaced all cycles of length 1 by contractible loops, hence by 1-gons. This does not change the surface hence it does not affect the genus of the embedding. In particular, if the embedding is minimal, the new embedding remains minimal (hence cellular), with the desired properties. Now by Lemma 2 each 1-gon consists of one bifacial edge. $\blacksquare$

9 Proof of the Genus Formula

9.1 Preliminary results

Consider a minimal embedding of an automaton $\mathtt{A}$ into a closed oriented surface Σ. We let e_0 denote the number of 0-cells (points, i.e. states), e_1 the number of 1-cells (open transitions) and e_2 the number of 2-cells (that is, the number of connected components of $\Sigma - \mathtt{A}$). The first classical result is Euler's formula ([Eul36], [Bre93, Chap. IV, §13]) that relates the genus to the CW-decomposition of Σ. In our context, since Σ is oriented and minimal, the formula takes the following form.

Lemma 5 (Euler's formula).

$$\chi(\Sigma) = 2 - 2g(\mathtt{A}) = e_0 - e_1 + e_2. \tag{7}$$

Another useful observation is a consequence of the decomposition $\pi_0(\Sigma - \mathtt{A}) = \coprod_{k \geq 0} F_k$. Namely,

$$e_2 = f_1 + f_2 + f_3 + \ldots = \sum_{k \geq 0} f_k. \tag{8}$$

The sum above is finite since the total number of 2-cells is finite. In particular, there is a maximal index $k \geq 0$ such that $f_k > 0$ and $f_l = 0$ for all $l > k$.

We need one more result that relates the number of 1-cells to the number of faces.

Lemma 6.
$$2e_1 = f_1 + 2f_2 + 3f_3 + \ldots = \sum_{k \geq 0} k \ f_k. \tag{9}$$

Proof. We begin with the relation (1): $\sum_{c\in\pi_0(\Sigma-\mathbf{A})}\langle e,c\rangle = 2$. It follows that

$$\sum_{e\in\Sigma^1}\sum_{c\in\pi_0(\Sigma-\mathbf{A})}\langle e,c\rangle = 2|\Sigma^1| = 2e_1.$$

Now use the decomposition of the cells into k-faces: $\pi_0(\Sigma-\mathbf{A}) = \coprod_{k\geq 0} F_k$.

$$\sum_{e\in\Sigma^1}\sum_{c\in\pi_0(\Sigma-\mathbf{A})}\langle e,c\rangle = \sum_{c\in\pi_0(\Sigma-\mathbf{A})}\sum_{e\in\Sigma^1}\langle e,c\rangle = \sum_{k\geq 0}\sum_{c\in F_k}\sum_{e\in\Sigma^1}\langle e,c\rangle$$

$$= \sum_{k\geq 0}\sum_{c\in F_k} k$$

$$= \sum_{k\geq 0} k\ f_k,$$

where we used the relation $\sum_{e\in\Sigma^1}\langle e,c\rangle = k$ for a k-face c. This completes the proof. ∎

9.2 Proof of Theorem 5 (Genus formula)

Consider a cellular embedding of $\mathbf{A}$ into a closed oriented surface Σ. Euler's formula (7) for the genus of Σ gives $g_\Sigma = 1 - \frac{e_0 - e_1 + e_2}{2}$. Since the automaton is complete, each state has exactly m outgoing transitions. Therefore $e_0 = e_1/m$. Next use the relations (9) and (8) to express e_1 and e_2 in terms of the k-faces. This yields the formula

$$g_\Sigma = 1 + \sum_{k=1}^{+\infty} \frac{k(m-1) - 2m}{4m} f_k.$$

Now $g(\mathbf{A}) \leq g_\Sigma$ with equality if and only if the embedding into Σ is minimal. This achieves the proof. ∎

10 Proof of the Genus Growth Theorem

It is convenient to introduce the following functions:

$$A(n) = \sum_{k\geq 3} \frac{k(m-1) - 2m}{4m} f_k(n) \text{ and } B(n) = \sum_{k\geq 3} k\ f_k(n).$$

We begin with

Lemma 7. *There is a constant $\alpha > 0$ such that*

$$A(n) \geq \alpha B(n). \tag{10}$$

Proof. To prove the claim, we first find $\alpha > 0$ such that

$$\frac{(m-1)k - 2m}{4m} \geq \alpha \, k \quad \text{for all } k \geq 3.$$

It suffices, therefore, to choose α such that

$$\frac{m-1}{4m} - \frac{1}{2k} \geq \alpha \quad \text{for all } k \geq 3.$$

This condition is satisfied if we choose

$$\inf_{k \geq 3} \left(\frac{m-1}{4m} - \frac{1}{2k} \right) = \frac{m-3}{12m} = \alpha_0 \geq \alpha.$$

(Note that $\alpha_0 > 0$ for $m \geq 4$.) This proves the lemma. ∎

Lemma 8.

$$\lim_{n \to +\infty} \frac{f_j(n)}{B(n)} = 0 \quad \text{for } j = 1, 2. \tag{11}$$

Proof. Since $f_k \leq z_k$, we have $\frac{f_k}{n} \leq \frac{z_k}{n} \underset{n \to +\infty}{\to} 0$ for $k = 1, 2$. Thus for any positive constants a, b,

$$\frac{n}{af_1 + bf_2} \underset{n \to +\infty}{\to} +\infty. \tag{12}$$

Observe that $2 \, mn = e_1(n) = f_1(n) + 2f_2(n) + B(n)$. Hence

$$n = \frac{1}{2m}(f_1(n) + 2f_2(n) + B(n)).$$

Replacing n in (12) by this expression, with $a = 1/(2m)$ and $b = 1/m$, we find that

$$\frac{\frac{1}{2m}B(n)}{\frac{1}{2m}f_1(n) + \frac{1}{m}f_2(m)} = \frac{B(n)}{f_1(n) + 2f_2(n)} \underset{n \to +\infty}{\to} +\infty.$$

Then

$$\max\left(\frac{B(n)}{f_1(n)}, \frac{B(n)}{f_2(n)} \right) \geq \frac{B(n)}{f_1(n) + 2f_2(n)} \underset{n \to +\infty}{\to} \infty,$$

as desired. ∎

Let us come to the proof of the Genus Growth Theorem. Let $\alpha > \varepsilon > 0$ satisfying the condition of Lemma 7. Lemma 8 ensures there is $N > 0$ such that for any $n \geq N$,

$$\left(\alpha + \frac{m+1}{4m} \right) f_1(n) + \left(2\alpha + \frac{1}{2m} \right) f_2(n) \leq \varepsilon \, B(n).$$

Hence for $n \geq N$,

$$A(n) \geq (\alpha - \varepsilon)B(n) + \left(\alpha + \frac{m+1}{4m}\right) f_1(n) + \left(2\alpha + \frac{1}{2m}\right) f_2(n).$$

Thus

$$A(n) - \frac{m+1}{4m} f_1(n) - \frac{1}{2m} f_2(n) \geq (\alpha - \varepsilon)(B(n) + f_1(n) + 2f_2(n))$$
$$= 2(\alpha - \varepsilon)e_1(n)$$
$$= 2(\alpha - \varepsilon)mn.$$

According to Theorem 5, $A(n) - \frac{m+1}{4m} f_1(n) - \frac{1}{2m} f_2(n) = g(n) - 1$. Thus

$$g(n) \geq 1 + 2(\alpha - \varepsilon)m \, n.$$

This achieves the proof of the theorem. ■

11   Conclusion

The topological tool we employ here, the genus as a complexity measure of the language, leads to a viewpoint that seems orthogonal to the standard one: it is not compatible with set-theoretic minimization (that is, state minimization). However, the genus does behave similarly to the state complexity with respect to operations such as determinization and union (up to a linear factor); furthermore, there is a hierarchy of regular languages based on the genus. This suggests a more systematic study of all operations : e.g. concatenation, star-operation, and composition of those. We take up this task in a sequel to this paper [BD13].

References

BD13.   G. Bonfante, F. Deloup, *The genus of a regular language II*, in preparation.

BP99.   I. Bezáková, M. Pál, *Planar Finite Automata*, Comenius University, Bratislava (1999).

BoCh76.   R.V. Book, A. K. Chandra, *Inherently Nonplanar Automata*, Acta informatica **6** (1976), 89–94.

Bre93.   G. E. Bredon, *Geometry and topology*, Graduate Texts in Math., no. 139, Springer, New York, 1993.

Car09.   A. Carbone, *Logical structures and genus of proofs*, Annals of Pure and Applied Logic **161** (2009), 139–149.

CKC83.   R.-W. Chen, Y. Kajitani, and S.-P. Chan, *A graph-theoretic via minimization algorithm for two-layer printed circuit boards*, IEEE transactions on circuits and systems **30** (1983).

Eul36.   L. Euler, *Solutio problematis ad geometriam situs pertinentis*, Commentarii Academiae Scientiarum Imperialis Petropolitanae **8** (1736), 128–140, `http://math.dartmouth.edu/~euler/docs/originals/E053.pdf`. (Reprinted and translated in Biggs, N. L.; Lloyd, E. K.; Wilson, R. J. (1976), Graph Theory 1736-1936, Oxford University Press.)

GMRY12.   Y. Gao, N. Moreira, R. Reis, and S. Yu, *A review on state complexity of individual operations*, Tech. report, Universidade do Porto, 2012.

Jae85.   F. Jaeger, *A survey of the cycle double cover conjecture*, Annals of Discrete Mathematics **27** in Cycles in Graphs, North-Holland Mathematics Studies **27** (1985), 1–12.

HMU06.   J. Hopcroft, R. Motwani, and J. Ullman, *Introduction to automata theory, languages, and computation (3rd edition)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

Myh57.   J. Myhill, *Finite automata and the representation of events*, Tech. report, WADD TR-57-624, Wright Patterson AFB, Ohio, 1957.

Ner58.   A. Nerode, *Linear automaton transformations*, Proceedings of the American Mathematical Society **9** (1958), 541 – 544.

RS97.   G. Rozenberg and A. Salomaa (eds.), *Handbook of formal languages*, vol. 1, Springer, 1997.

Sak09.   J. Sakarovitch, *Elements of automata theory*, Cambridge University Press, 2009.

SHL90.   M. Stallmann, T. Hughes, and W. Liu, *Unconstrained via minimization for topological multilayer routing*, IEEE Transactions on computer-aided design **9** (1990), no. 9.

Sta74.   R. Statman, *Structural complexity of proofs*, Ph.D. dissertation. Department of Philosophy, Stanford University, 1974.

Xuo77.   N. Huy Xuong, *Sur quelques problems d'immersion d'un graphe dans une surface*, Ch. 3, Ph.D. Thesis (1977), Grenoble, France.

You63.   J.W.T. Youngs, *Minimal imbeddings and the genus of a graph*, J. Math. Mech. **12** (1963), 303–305.

Yu00.   S. Yu, *State complexity of regular languages*, Journal of Automata, Languages and Combinatorics **6** (2000), 221–234.

G. Bonfante, LORIA, Université de Lorraine, Campus scientifique, B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex

    `guillaume.bonfante@loria.fr`

F. Deloup, IMT, Université Paul Sabatier – Toulouse III, 118 Route de Narbonne, 31069 Toulouse cedex 9

    `florian.deloup@math.univ-toulouse.fr`